

Логические методы классификации

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

Спецкурс «Теория статистических решений».
Лекция 5.

Общая характеристика

Логические методы — широко используемый класс методов.

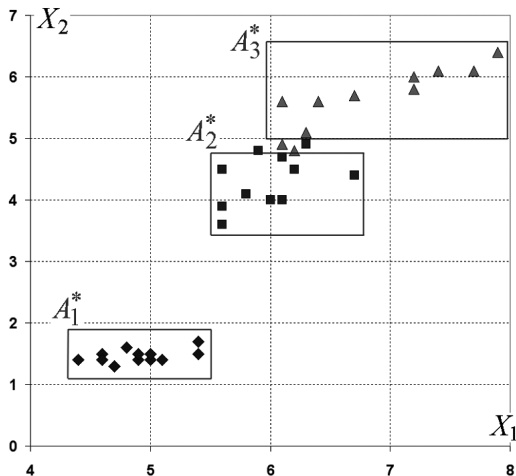
Основные варианты:

- решающие списки,
- решающие деревья.

Свойства:

- работа в разнотипном пространстве,
- работа с пропусками,
- интерпретируемость решений.

Логические закономерности для задачи Iris



Понятие закономерности

X – пространство значений прогнозирующих переменных,
 $Y = \{0, 1, \dots\}$ – прогнозируемая переменная, $\varphi : X \rightarrow \{0, 1\}$ – предикат.

$V = \{(x^i, y^i) \mid i = \overline{1, N}\}$ – выборка объектов,

M – из них 1-го класса,

n – число точек, на которых предикат истинный,

m – из них 1-го класса.

«Хороший» предикат — закономерность:

$$a = \frac{m}{M} \rightarrow \max, \quad b = \frac{n - m}{n} \rightarrow \min.$$

Статистический критерий

Вероятность при отборе n объектов получить m из них 1-го класса:

$$P(m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = \frac{C_n^m C_{N-n}^{M-m}}{C_N^M}.$$

Критерий «неслучайности» отбора:

$$P(m \geq m_0) = \sum_{m=m_0}^M P(m) < \alpha.$$

Информационный критерий

Формула Стирлинга:

$$\ln(k!) \approx k \ln k - k + \frac{1}{2} \ln(2k\pi) + \frac{1}{12k} - \frac{1}{360k^2}.$$

Количество информации:

$$G = H\left(\frac{M}{N}\right) - \frac{n}{N} \cdot H\left(\frac{m}{n}\right) - \frac{N-n}{N} \cdot H\left(\frac{M-m}{N-n}\right).$$

$$H(p) = -p \ln p - (1-p) \ln(1-p), \quad G \approx -\frac{1}{N} \ln P(m).$$

Принцип равномерной сходимости

Статистический критерий $P(m \geq m_0) < \alpha$ характеризует «случайность» только для априорно выбранной закономерности.

Поскольку возможных предикатов много, вероятность того, что хотя бы на одном значении критерия будет «хорошим», больше.

Вероятность зависит от сложности предиката.

Поиск закономерностей

- Бинаризация признаков.
- Интервальные предикаты.
- Конъюнкции элементарных предикатов.

Алгоритмы поиска

- КОРА.
- ТЭМП.

Решающие списки

Метод классификации

- Формируем упорядоченный по информативности список закономерностей.
- Решение принимаем по первой закономерности, которой удовлетворяет объект.

Можно применять голосование.

Критерии ветвления

- Число ошибок.
- Информационный.
- Гини.
- Число пар объектов разных классов, разделяемых предикатом.

Алгоритмы

- Жадный.
- Рекурсивный.
- Неограниченный.

Недостатки логических методов

- Невозможность «гладких» решений.
- Много эвристик.
- Вычислительная трудоёмкость нахождения точных решений.
- Выбор оптимальных критериев.